https://doi.org/10.3311/PPci.12559
Creative Commons Attribution

Periodica Polytechnica Civil Engineering, 63(1), pp. 87–92, 2019

Application of MARSplines Method for Failure Rate Prediction

Małgorzata Kutyłowska[1*]

[1] Faculty of Environmental Engineering
Wrocław University of Science and Technology
Wybrzeże St. Wyspiańskiego 27, 50-370 Wrocław, Poland
* Corresponding author, e-mail: malgorzata.kutylowska@pwr.edu.pl

Received: 21 May 2018, Accepted: 10 October 2018, Published online: 06 November 2018

Abstract

In this paper MARSplines method was presented to model failure rate of water pipes in years 2015-2016 in the selected Polish city. The output parameters were chosen as three dependent variables - three values of failure rate of water mains, distribution pipes and house connections. Diameter, season, material and kind of the conduit were selected as independent variables. At the beginning of modelling 21 basis (splines) function were assumed. On a final note two functions were selected (after reduction of negligible functions). The model consists of three factors: β_0, β_1 and β_2. The penalty for adding basis function was assumed at the level of 2. The correlation was equalled to 0.44. Relatively huge discrepancies between real and predicted values of failure rate of water mains and house connections were observed. In the future investigations concerning this problem the three separated models for each kind of conduit should be created. The calculations using MARSplines method were carried out in the program Statistica 13.1.

Keywords

failure analysis, predictors, regression method, water pipes

1 Introduction

Water distribution systems are now commonly used almost everywhere in Poland. The statistics for 2015 [1] show that 96.5% of the population is connected to the water-pipe network. In 2015 the total length of the water-pipe network in rural areas amounted to 240000 km. In comparison with the year 2005, network length has increased by about 50000 km [1]. In this context the emphasis as regards water supply systems should be put on the use of new approaches of estimating of water losses and pipes' failures [2] and on the chemical stability of water delivered to the consumers [3] as well as modelling of water consumption [4]. Considering the current research on this subject in Poland [5], [6], [7], [8], [9], [10] and abroad [11], [12], based on operational data and on laboratory studies of the frequency of failures of water pipes, it seems that such research needs to embrace new methods of predicting reliability indices, i.a. the failure rate. In recent years predictive methods, such as the support vector method [13], the K-nearest neighbours method [14], the regression and classification trees method [15] and artificial neural networks [16], have become popular in the modelling of various broadly understood engineering problems. A method exploiting the "splining" of many functions, called the MARS (Multivariate Adaptive Regression Splines) method, is one of the regression algorithms which can be applied to solve variables prediction problems not easily described by typical mathematical models [17].

The primary aim of this paper is to show the potential of MARSplines nonparametric regression method for describing the failure frequency of water pipes. The available publications on the subject indicate that currently this algorithm is not widely used to solve engineering problems. Recently it was used to locate places with the highest risk of landslides [18] and to predict the effect of changes in temperature on the formation of crystals [19]. Since the MARS method has not been applied to the modelling of the technical condition of underground facilities the author decided to try this out. Moreover, in the course of further research it will be possible to compare several regression methods and choose the optimal algorithm for predicting the failure rate of water conduits in selected Polish towns.

2 Material and methods

The failure rate (λ, fail./(km·year)) of water mains – WM (λm), distribution pipes – DP (λr) and house connection – HC (λp), in a selected seaside town with a population of

Cite this article as: Kutyłowska, M. "Application of MARSplines Method for Failure Rate Prediction", Periodica Polytechnica Civil Engineering, 63(1), pp. 87–92, 2019. https://doi.org/10.3311/PPci.12559

over 41000 was the dependent variables predicted using the MARSplines method. Operational data for the years 2015-2016, obtained from a water company, were used for modelling. The training sample and the testing sample, randomly generated from the whole data set, amounted to respectively 70% (66 cases) and 30% (26 cases). Water pipe diameter was the quantitative independent variable while material and type of conduit were the qualitative independent variables. The non-heating (summer) season was assumed to last from the beginning of March to the end of October. In the years 2015–2016 on average the values of failure rate λ calculated on the basis of the operational data amounted to: 0.10, 0.24 and 0.30 for the water mains, the distribution pipes and the house connections, respectively. For the heating (winter) season (H) and the non-heating season (NH) the failure rate amounted to: λmh = 0.18 and λmnh = 0.06, λrh = 0.23 and λrnh = 0.25, λph = 0.33 and λpnh = 0.29 fail./(km·year). In the system of water mains all the conduits with a diameter larger than 300 mm were considered. The diameter of the distribution pipes ranged from 100 to 300 mm. The house connections were pipelines with a diameter of up to 90 mm inclusive. The water conduits in the analysed system are made of asbestos cement (AC), PE, steel (S) and cast iron (CI). The house connections are made of steel and PE while the other types of conduits (water mains and distribution pipes) are made of all the materials mentioned above.

According to the best knowledge the problem of the reliability level, technical condition and failure rate of water-pipe network was not solved till now using MARSplines methodology. For modelling purposes very basic exploitation data were used. The first step in prediction by means of nonparametric method should be to check if at all proposed methodology (in this case MARSplines algorithm) is useful for estimation of failure frequency and reliability level. In this connection at the beginning, data listed in the table 1 were treated as independent and dependent variables. In the next step of investigation should be including more precise information about water-pipe network and its exploitation.

MARSplines is a multivariate adaptive regression which uses spline functions (curves) [20]. It can be applied to solve both classification problems, i.e. problems in which the dependent (predicted) variable is a qualitative variable, and regression problems where the predicted value is a random variable. In both cases the independent variables (the predictors) can be quantitative and qualitative variables. An important advantage of the MARSplines method (which is a nonparametric algorithm) is that one does not need to know the functional relation between the dependent variables and the independent variables. It is not necessary to a *priori* define the type of function (e.g. a linear, logistic, etc., function) as in other predictive methods, such as SVM and artificial neural networks. A dependence between the two kinds of variables is found solely through an analysis of the set of coefficients and basis functions acquired from the modelling data. The MARSplines method can be described as a segmented and multiple linear regression. The applicability of the linear functions is closely bound with the boundaries of the segments. Moreover, the input data collected in the decision space are divided to obtain separate subsets with appropriate regression or classification functions. This is an advantage mainly in the case of an extensive (comprising many variables) vector of input variables. In such a case, other predictive algorithms can encounter the multidimensionality problem. The above mentioned coefficients defining the influence of a given predictor on the dependent variable can be compared with one another (for different independent variables) only when the variables are normalized to a zero average and a unit standard deviation. Otherwise, when the predictors are measured on different scales, this approach has no mathematical sense. In the MARSplines method the basis functions are linear functions (t-x) and (x-t). Parameter t, the value of which depends on the problem currently being solved and on the analysed data, is a basis function node. Using the basis functions and the model parameters (determined by the least squares method) one predicts the dependent variable on the basis of the input data. The general equation in the MARSplines method can be written as follows [20]:

$$y = f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X) \tag{1}$$

Addition is performed for all the K functional model components. Dependent variable y is calculated as a function of independent variables X (and their interactions). The elements of this function are: the initial ordinate (β_0) and the weighted (with weights β_k) sum of one or many basis functions h_k(X). In other words, this model is the sum of the basis functions selected (from many available functions) to solve a specific problem.

The primary truncated basis functions for modelling the dependent variable are expressed as follows [20]:

$$(x-t) = \begin{cases} x-t, & x > t \\ 0, & x \le t \end{cases} \tag{2}$$

Table 1 Ranges of predictors

Season	Material	Type	Diameter, mm
Training sample			
H, NH	AC, PE, S, CI	WM, DP, HC	15–500
Testing sample			
H, NH	AC, PE, S, CI	DP, HC	32–250

Table 2 Example of matrix of learning sample

Season	Material	Type	Diameter	Failure rate
winter	S	DP	100	0.29
winter	AC	DP	100	0.29
winter	AC	DP	100	0.29
winter	CI	DP	100	0.29
summer	S	DP	100	0.29
summer	S	DP	100	0.29
summer	S	DP	100	0.29
summer	S	HC	50	0.50
summer	PE	HC	40	0.50
summer	CI	WM	500	0.06

The MARSplines algorithm is used to search the space of all the input data values and to analyse the interactions between the data. The aim is to maximize the goodness of fit and to find the most vital predictors. There is a risk of overtraining on the training data set (then the testing data will not be correctly predicted) since MARSplines as a nonparametric model exceptionally well fits to data. Trimming, i.e. the reduction of basis functions, is a way of avoiding too excessive goodness of fit of the predicted variable to the actual values. The reduction of the number of basis functions entails the selection of the most important predictors. Functions which have a significant bearing on the prediction of the dependent variable [20] and the ones the removal of which causes the smallest increase in the square error are selected. Calculations by the MARSplines method were performed using Statistica 13.1. After removing insignificant functions, maximally 21 basis functions were used. The order of interaction amounted to 1 while the penalty was equal to 2 and the threshold to 0.0005. The interaction order of 1 takes into account the main effects between the variables while the order of 2 takes into account all the interactions between the pairs of variables. A specified penalty is added at each instant when another basis function is added to the algorithm. The threshold prevents overtraining and excessive fitting of the model results to the actual results.

Table 1 shows the values of the independent variables used to build the MARSplines model and to predict the dependent variable – the failure rate of water conduits.

It should be noted that in the testing sample the range of two independent variables, i.e. type of conduit and diameter, was narrowed. This is due to the fact that the cases for the particular samples had been randomly selected by the same algorithm which is used to select subsets in the artificial neural network algorithm. In the present study the problem of water conduit failure rate modelling is considered for the heating season and the non-heating season. Previously this division was not used in the author's investigations of the problem of failure rate prediction by regression methods and the failure frequency level was modelled for a given selected period (usually a few years, without the division into winter and summer months) [21].

Table 2 shows the extract (10 cases) from the whole matrix (66 cases) of learning data. The matrix consists of independent variables and output data (failure rate). The model was built and learnt using the whole matrix and special toolbox (just MARSplines) in Statistica software was adopted to forecast failure frequency. All data were registered by water utility and are real one. Similar matrix was created for testing the model using testing sample of data.

3 Results and discussion

In the MARSplines method the quality of the model is expressed by the GCV (*Generalized Cross Validation*) coefficient which is an approximate cross validation coefficient and is calculated from the relation [17, 20, 22]:

$$GCV = \frac{\sum_{i=1}^{N}\left(y_i - f\left(x_i\right)\right)^2}{N\left(1-\frac{C}{N}\right)^2} \tag{3}$$

$$C = 1 + c \cdot d \tag{4}$$

N – number of cases,
C – penalty for adding the next basis function,
d – effective of degrees of freedom; equals to the number of independent functions,
c – controls the penalty value.

The model describing the failure rates of the water mains, the distribution pipes and the house connections for the respective seasons was characterized by GCV = 0.010508. Two basis functions (the number of functions was reduced by as many as 19) and three factors, i.e. a free term (β_0 = 0.281709) and two coefficients (β_1 = 0.134385 and β_2 = –0.000296), were sufficient to solve the failure frequency problem. Such variables as diameter and type of conduit (only house connection) were taken into account

in a failure rate modelling relation. One of features of the MARSplines method is the generation of an equation describing the considered problem, which includes the two factors mentioned above. The frequency of failures of the water pipes in the considered town can be described by the following relation:

$$\lambda = 0.281709 + 0.134385 \cdot max(0;\ type\_HC - 0) - 0.000296 \cdot max(0;\ diameter - 15) \quad (5)$$

It is somewhat surprising and inexplicable why "type_HC" in particular was the predictor used to create (by splining two basis functions) the global function describing the failure frequency of all the types of pipelines. The number of training cases for the house connections was not dominant at all. From the 66 training cases 5 concerned water mains, 37 – distribution pipes and 24 – house connections. Perhaps this is due to the MARSplines algorithm's non-parametricity. If at the start no functional relation between the predictors and the dependent variables is assumed, then the type of function (Equation (5)) is acceptable. But one should also pay attention to the correlation between the actual values and the predicted ones. Determination coefficient $R^2 = 0.44$ is not a satisfactory value and indicates discrepancies between the model results and the operational failure rate values. The discrepancies are also observed in Figs. 1–3 which show the actual (A) and predicted (P) values of failure rates λm, λr and λp. Symbols H and NH stand for heating season and non-heating season, respectively. The results presented below are for both the training set and the testing set since the modelling results are identical in both cases, which indicates, on one hand, that the model was not overtrained and, on the other hand, that further research on the MARSplines method with regard to the modelling of the frequency of failures of water conduits is needed since the results obtained so far are not satisfactory from the engineering point of view.

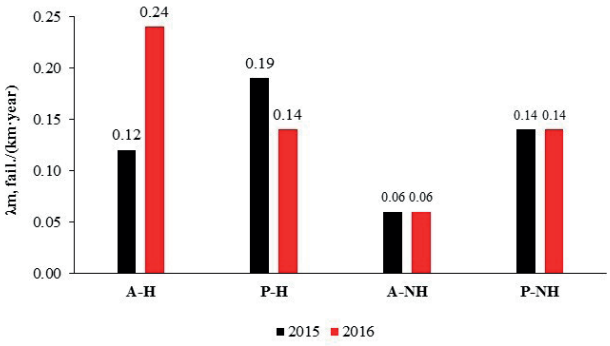

Fig. 1 Real and predicted values of failure rate of water mains

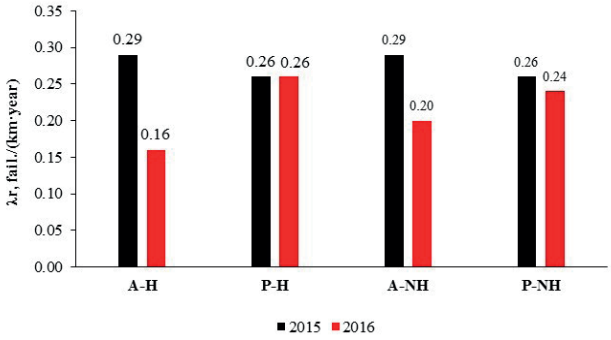

Fig. 2 Real and predicted values of failure rate of distribution pipes

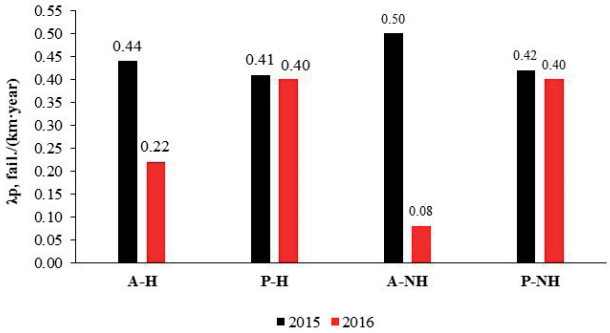

Fig. 3 Real and predicted values of failure rate of house connections

An analysis of Fig. 1 shows that the failure rate of the water mains was not correctly modelled by the MARSplines method. In the year 2015 (both the heating season and the non-heating season) and in the non-heating season of 2016 the λm values are significantly overestimated. In the 2016 heating season the failure rate is underestimated. The water mains perform a significant and the most important role in the whole water distribution system. A failure of a conduit of this type sometimes results in no water supply to a considerable area. Therefore the failure rate for water mains should be predicted most accurately since an incorrect prediction can lead to disastrous consequences, e.g. an incorrect selection of conduits for renovation or replacement can result in a serious construction failure.

Indicator λr was modelled a little better. The differences between the actual values and the predicted ones are insignificant. Similarly as in the case of water mains, the estimate for the heating season was inaccurate. For distribution pipes the MARSplines algorithm was found to be a relatively good method of modelling the failure rate. However, in the considered case there were three dependent variables (λm, λr, λp) at the model's output whereby the model should be considered as a whole, i.e. for all the three types of conduits. Thus, even though the modelling results for the distribution pipes are acceptable, the lack of correlation for the water mains and the house connections is conspicuous and should be corrected.

However, Fig. 3 shows that the modelled failure frequency level for the house connections differs considerably from the operational values. Only for the heating season of 2015 the correlation between the predicted values and the actual ones is acceptable. Perhaps in the future analyses one should consider modelling each of the failure rates separately, i.e. creating separate models for describing the failure rate of respectively water mains, distribution pipes and house connections, which has already been done in the case of other regression methods [21].

4 Conclusions

The obtained prediction results indicate that even though the MARSplines method has many advantages it is not an algorithm which could be unreservedly used to analyse the failure rate of water conduits. The main limitation is using linear relation between dependent and independent variables. As far as it is known the relationships between operational parameters as e.g. material, diameter and season are not always described by linear function. Extending the model by adding new independent variables is also limited just by the fact that MARSplines can be described as a segmented and multiple linear regression. In many cases such limitation to only linear regression could be a problem resulting in improper prediction of dependent variable. A different approach to model building seems to be necessary, i.e. separate models for each type of conduit should be created, which can lead to more accurate prediction results. Moreover, perhaps the range of the data included in the analysis should be widened so that even random sampling will generate data sets comprising the same predictors, but with a different range of assumed values (tab. 1). It is also necessary to check this nonparametric method by applying it to other water supply systems, whereby it will be possible to draw further conclusions and compare this kind of modelling with other predictive algorithms. The obtained results of modelling (not only by means of MARSplines method, but also by other regression algorithms) could be useful for water utility to predict failure frequency, plan the modernization schedule and to get to know what kind of operational and exploitation data should be registered. Sometimes the range and scope of information which is collected by water utility is not enough for scientific purposes and such investigations as e.g. modelling can be helpful for better understanding the forecasting problem and for minimizing the risk of damage by proper management [23].

Acknowledgement

This research has been carried out as part of the statutory activity of the Faculty of Environmental Engineering at Wrocław University of Science and Technology, funded by the Ministry of Science and Higher Education in the years 2018-2019, project no. 0401/0054/18.

References

[1] Statistical Yearbook of the Republic of Poland, Central Statistic Office, Warsaw, 2016.

[2] Suchorab, P., Kowalska, B., Kowalski D. "Numerical investigations of water outflow after the water pipe breakage", Annual Set The Environment Protection, 18(2), pp. 416–427, 2016. Available at http://ros.edu.pl/images/roczniki/2016/No2/31_ROS_N2_V18_R2016.pdf [Accessed: 06.11.2018]

[3] Pietrucha-Urbanik, K., Tchórzewska-Cieślak, B., Papciak, D., Skrzypczak, I. "Analysis of chemical stability of tap water in terms of required level of technological safety", Archives of Environmental Protection, 43(4), pp. 3–12, 2017. https://doi.org/10.1515/aep-2017-0043

[4] Piasecki, A., Jurasz, J., Kaźmierczak, B. "Forecasting Daily Water Consumption: a Case Study in Torun, Poland", Periodica Polytechnica Civil Engineering, 62(3), pp. 818–824, 2018. https://doi.org/10.3311/PPci.11930

[5] Pietrucha-Urbanik, K., Studziński, A. "Case study of failure simulation of pipelines conducted in chosen water supply system", Eksploatacja i Niezawodnosc – Maintenance and Reliability, 19(3), pp. 317–323, 2017. https://doi.org/10.17531/ein.2017.3.1

[6] Iwanek, M., Suchorab, P., Karpińska-Kiełbasa, M. "Suffosion holes as the result of a breakage of a buried water pipe", Periodica Polytechnica Civil Engineering, 61(4), pp. 700–705. 2017. https://doi.org/10.3311/PPci.9728

[7] Boryczko, K., Tchórzewska-Cieślak, B. "Analysis of risk of failure in water main pipe network and of delivering poor quality water", Environment Protection Engineering, 40(4), pp. 77–92, 2014. https://doi.org/10.5277/epe140407

[8] Piegdoń, I., Tchórzewska-Cieślak, B., Eid, M. "Managing the risk of failure of the water supply network using the mass service system", Eksploatacja i Niezawodnosc – Maintenance and Reliability, 20(2), pp. 284–291, 2018. https://doi.org/10.17531/ein.2018.2.15

[9] Iwanek, M., Kowalski, D., Kwietniewski, M. "Model studies of a water outflow from an underground pipeline upon its failure", Ochrona Środowiska, 37(4), pp. 13–17, 2015. Available at http://www.os.not.pl/docs/czasopismo/2015/4-2015/Iwanek_4-2015.pdf [Accessed: 06.11.2018] (In Polish)

[10] Musz-Pomorska, A., Iwanek, M., Parafian, K., Wójcik, K. "Analysis of water losses in two selected water distribution systems". In: 9th Conference on Interdisciplinary Problems in Environmental Protection and Engineering EKO-DOK, Boguszow-Gorce, Poland, 2017. pp. 468–475. https://doi.org/10.1051/e3sconf/20171700062

[11] Wilson, D., Moore, I., Filion, Y. "Using sensitivity analysis to identify the critical factors that lower the factor of safety of large-diameter cast iron mains", Urban Water Journal, 14(7), pp. 685–693, 2017. https://doi.org/10.1080/1573062X.2016.1236137

[12] Ciaponi, C., Franchioli, L., Papiri, S. "Simplified procedure for water distribution networks reliability assessment", Journal of Water Resurces Planning and Management, 138(4), pp. 368–376, 2012. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000184

[13] Rusek, J. "Application of Support Vector Machine in the analysis of the technical state of development in the LGOM mining area", Eksploatacja i Niezawodnosc – Maintenance and Reliability, 19(1), pp. 54–61, 2017. https://doi.org/10.17531/ein.2017.1.8

[14] Meng, Q., Cieszewski, Ch. J., Madden, M., Borders B. E. "K nearest neighbor method for forest inventory using remote sensing data", GIScience and Remote Sensing, 44(2), pp. 149–165, 2007. https://doi.org/10.2747/1548-1603.44.2.149

[15] Irimia-Dieguez, A. I., Blanco-Oliver, A., Vazquez-Cueto, M. J. "A comparison of classification/regression trees and logistic regression in failure models", Procedia Economics and Finance, 23, pp. 9–14, 2015. https://doi.org/10.1016/S2212-5671(15)00493-1

[16] Kutyłowska, M. "Prediction of failure frequency of water-pipe network in the selected city", Periodica Polytechnica Civil Engineering, 61(3), pp. 548–553, 2017. https://doi.org/10.3311/PPci.9997

[17] Kaveh, A., Bakhshpoori, T., Hamze-Ziabari, S. M. "M5' and Mars based prediction models for properties of self-compacting concrete containing fly ash", Periodica Polytechnica Civil Engineering, 62(2), pp. 281–294, 2018. https://doi.org/10.3311/PPci.10799

[18] Wang, L. J., Guo, M., Sawada, K., Lin, J., Zhang, J. "Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models", Catena, 135, pp. 271–282, 2015. https://doi.org/10.1016/j.catena.2015.08.007

[19] Antanasijević, J., Pocajt, V., Antanasijević, D., Trišović, N., Fodor-Csorba, K. "Prediction of clearing temperatures of bent-core liquid crystals using decision trees and multivariate adaptive regression splines", Liquid Crystals, 43(8), pp. 1028–1037, 2016. https://doi.org/10.1080/02678292.2016.1155769

[20] Statistica 13.1, Electronic Manual

[21] Kutyłowska, M. "Prediction of water conduits failure rate – comparison of support vector machine and neural network", Ecological Chemistry and Engineering A, 23(2), pp. 147–160, 2016. https://doi.org/10.2428/ecea.2016.23(2)11

[22] Materials from the workshop Statistica Statsoft "Data mining – prediction methods", Poland, Kraków, 2017.

[23] Zimoch, I. "Pressure control as part of risk management for a water-pipe network in service", Ochrona Środowiska, 34(4), pp. 57–62, 2012. Available at http://www.os.not.pl/docs/czasopismo/2012/4-2012/Zimoch_4-2012.pdf [Accessed:06.11.2018] (In Polish)